Linear Regression

Fundamentals, Methods, and Limitations

Badi Ahmed

Mathematics & Machine Learning Enthusiast

ahmedbadi905@gmail.com

December 3, 2025

Outline

Introduction to Linear Regression

What is Linear Regression?

Definition

A supervised learning algorithm modeling a continuous target with a linear equation

Key Concept

- Assumes linear relationship: $y = f(X)$
- Minimizes prediction errors
- Simple and interpretable

Practical Example

Predicting exam scores from study hours; goal: best-fit line

Why Linear Regression Matters

Six Key Advantages

1. Simplicity and clarity
2. Interpretability
3. Computational efficiency
4. Foundation for advanced models
5. Robust with preprocessing
6. Versatile extensions

Applications

- Housing prices
- Sales forecast
- Stock returns
- Crop yield
- Medical analysis

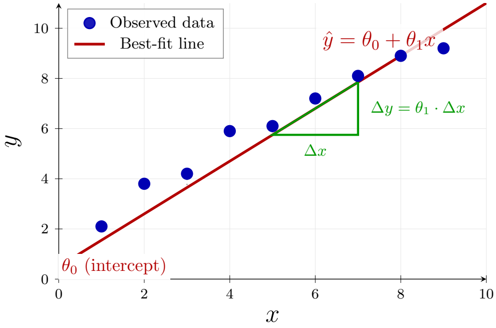
The Best-Fit Line

The Best-Fit Line Concept

Formulation

$$\hat{y} = \theta_0 + \theta_1 x$$

with θ_0 intercept and θ_1 slope.



Least Squares Method

Residual

$$e_i = y_i - \hat{y}_i = y_i - (\theta_0 + \theta_1 x_i)$$

SSE

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Key Insight

Least squares finds θ_0, θ_1 that minimize SSE.

Simple vs. Multiple Regression

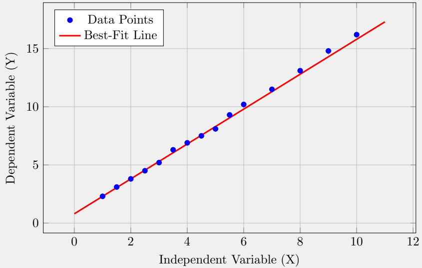
Simple:

$$\hat{y} = \theta_0 + \theta_1 x$$

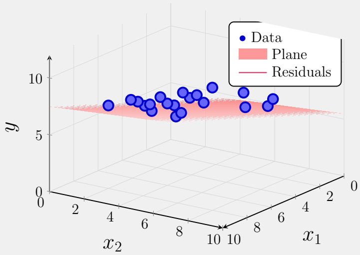
Multiple:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k$$

Figure



Multiple Regression: $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$



Hypothesis & Cost

Hypothesis Function

Simple Form

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Vector Form

$$h_\theta(x) = \theta^\top x = \sum_{j=0}^{k} \theta_j x_j, \quad x_0 = 1$$

Mean Squared Error

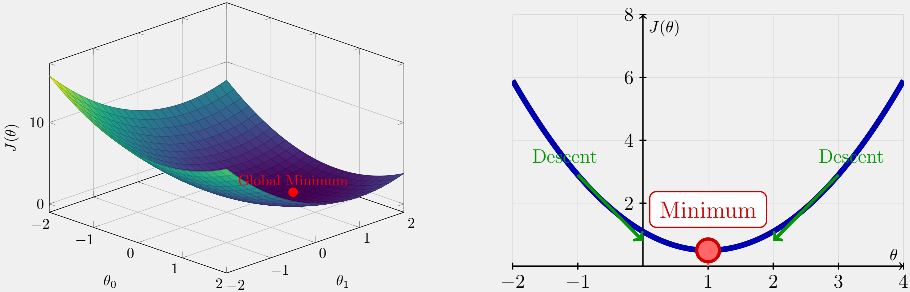
MSE:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x_i) - y_i)^2$$

For optimization:

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Figure



Optimization

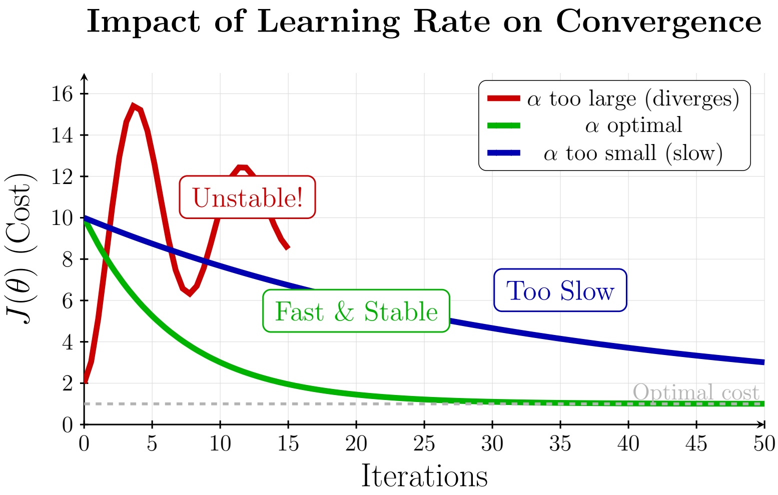
Gradient Descent

Update

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{n} \sum_{i=1}^{n} (h_\theta(x_i) - y_i), \quad \frac{\partial J}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} (h_\theta(x_i) - y_i) x_{i,j}$$

Choosing the Learning Rate

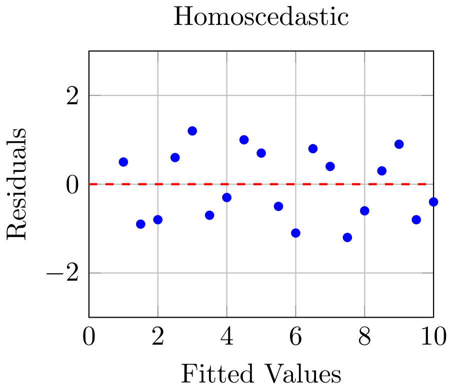


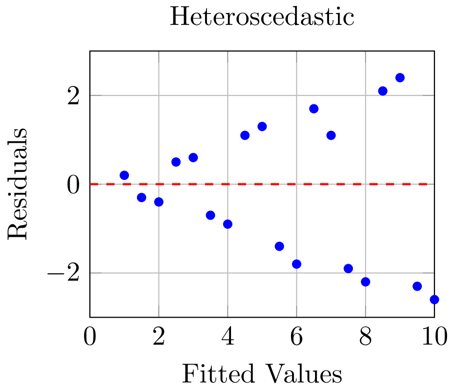
Assumptions

Six Key Assumptions

Linearity; Independence; Homoscedasticity

Normality; No Multicollinearity; Additivity





Evaluation Metrics

Core Metrics

$$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2, \quad \text{MAE} = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}, \quad R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Adjusted R^2 and Table

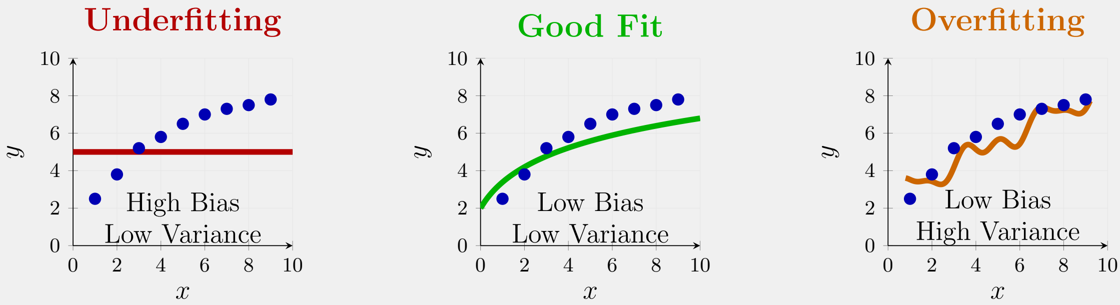
$$\text{Adj-}R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Metric	Unit	Sensitivity	Use Case
MAE	Target	Low (outliers)	Interpretability
RMSE	Target	High (outliers)	Benchmark
MSE	Squared	Very high	Optimization
R^2	0–1	Fit	Variance explained

Regularization

Overfitting and Underfitting Problem

Figure



Ridge (L2)

$$J(\theta) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{k}\theta_j^2$$

- Shrinks coefficients; keeps all features
- Good with multicollinearity

Lasso (L1)

$$J(\theta) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{k}|\theta_j|$$

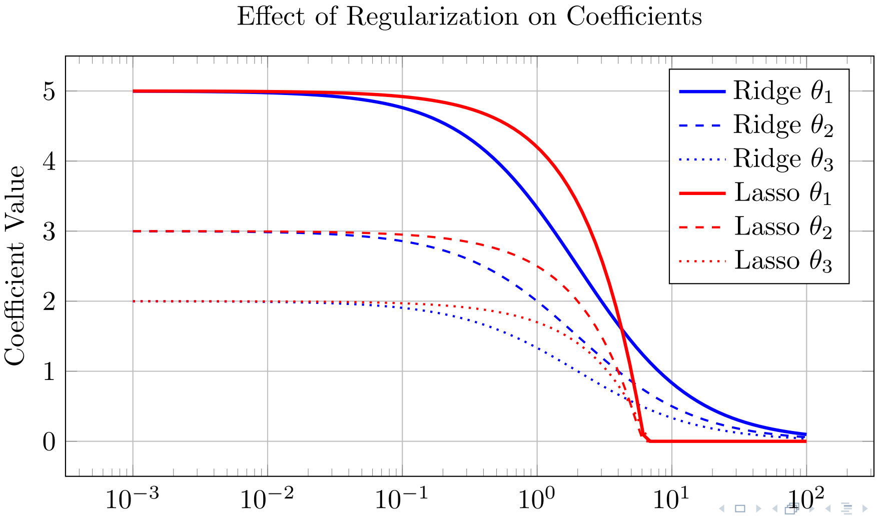
- Sets some coefficients to zero (feature selection)
- Sparse and interpretable

Elastic Net

$$J(\theta) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 + \lambda\left[\alpha\sum|\theta_j| + \frac{1-\alpha}{2}\sum\theta_j^2\right]$$

- Combines selection + shrinkage
- Stable with correlated features

Regularization



Advantages & Limitations

Advantages

- Simple, interpretable, efficient
- Strong theory; great baseline
- Versatile with extensions

Limitations

- Linearity assumption; outlier sensitivity
- Multicollinearity; independence assumption
- Extrapolation risk; limited pattern capacity

Conclusion

Key Takeaways

- Fundamentals and mechanics: least squares, gradient descent
- Assumptions govern applicability
- Regularization prevents overfitting

When to Use

Good fit: linear relation, interpretability, baseline

Poor fit: non-linear, many outliers, high-dim $p \gg n$

Thank You!

Questions & Discussion

Badi Ahmed

Linear Regression: Fundamentals, Methods, and Limitations

ahmedbadi905@gmail.com

December 3, 2025